Task-Oriented Slot-Based Cumulant Discovery in General Value Functions

Anonymous authors
Paper under double-blind review

Abstract

General value functions (GVFs) provide a reinforcement learning (RL) framework for learning predictive knowledge that can improve policy learning. Traditionally, the predictive questions that learned GVFs have to answer were hand-designed to be relevant to the task at hand. However, useful questions can also be automatically discovered by an auxiliary network that is trained via a meta-gradient approach which directly optimizes for the RL loss. Recent work has shown that using a slot-based representation as input to this question network facilitates the discovery of GVFs that make predictions about objects in environments with visual observations. In this work, we propose to leverage information only in the meta-gradient to steer discovery towards task-relevant features.

1 Introduction

General value functions (GVFs) (Sutton et al., 2011) can be used to derive *predictive knowledge* about the environment in a reinforcement learning (RL) setting. Akin to value functions that describe expected cumulative rewards in an Markov decision processes (MDP) setting, GVFs describe the expected cumulative sum of other scalar quantities than rewards. These quantities, or *cumulants* can either be hand-crafted (Sutton et al., 2011; Jaderberg et al., 2016) or learned in a goal-directed manner (Veeriah et al., 2019; Kearney et al., 2022; Nath et al., 2023). GVFs can be used as auxiliary tasks that facilitate representation learning for the main task (Veeriah et al., 2019) or they can directly be used as state inputs (Kearney et al., 2022). However, simultaneous learning of policies, cumulants and discount factors that together specify GVFs is a difficult optimization problem. If, however, we restrict the problem space to involve only on-policy GVFs, the problem of cumulant *discovery* becomes amenable, because we do not need any importance sampling. Nath et al. (2023) describe an approach of learning GVFs from visual observations to be used directly as state inputs. Their approach focuses on learning a limited number of GVFs, each associated with a single object. The set of objects is derived using slot attention (Locatello et al., 2020) and the overall architecture has been shown to work well.

However, this process is heavily dependent on the good separation of objects of slot attention, which may not always be the case in most domains. Additionally, slot attention can only distinguish objects by pixel-matching using a reconstruction loss, and as a result, objects that differ by any other topology will not be properly distinguished by slot attention. In our work, we remove separate slot attention training altogether, and only use the encoder of slot attention to capture task-specific features in the slots. We show that this **simpler** architecture performs on par with the architecture from Nath et al. (2023) that trains the slot attention module via reconstruction.

The main contributions of our paper are as follows:

- We demonstrate that adding discount prediction, Hungarian Loss and Huber TD losses to the overall architecture from Nath et al. (2023) improves its performance.

- Capturing object-centric cumulants is not essential for good performance in limited GVF regime; similar performance can be obtained even in the absence of the full slot attention framework, using just the meta-gradients from the main RL loss.

2 Background

Auxiliary tasks learned in parallel with the main RL task can aid representation learning by providing an additional learning signal for instance via modeling predictive knowledge. This can be useful particularly in sparse or noisy reward environments, and also prevent over-fitting to the current rewards.

General Value Functions (GVFs): A natural candidate for predicting auxiliary tasks are GVFs (Sutton et al., 2011). They are *value functions* which instead of discounted returns predict discounted sums of "cumulants" that are specified as a function of the environment state. Just like the usual value functions, GVFs can be learned using Temporal Difference (TD) Error. Given an MDP, the GVF for a particular state-action pair (s, a) would be:

$$Q^{GVF}(s, a; \pi, \gamma, c) = \mathbb{E}[G_t | s_t, a_t, A_{t+1:T-1} \sim \pi]$$

where, G_t is the discounted sum of cumulants with discounting factor γ, over a trajectory length T. The definition assumes that we are following a policy π and the returns are calculated based on the cumulants c, which can be arbitrary. Sutton et al. (2011) showed the utility of using GVFs to learn predictive knowledge about the environment.

Using GVFs as Auxiliary Tasks: Jaderberg et al. (2016) popularized the idea of using GVFs as auxiliary tasks as a way to improve representation learning in RL. The auxiliary tasks designed for that work were hand-crafted like pixel-control, feature-control based on changes in pixel intensities, and feature activations, respectively. However, to fully appreciate the utility of using GVFs as auxiliary tasks, we need to automatically discover good *cumulants* for these GVFs. Veeriah et al. (2019) developed an approach for automatically discovering such *cumulants*, which employs meta-gradients (Xu et al., 2018) to automatically learn task-specific cumulants driven by the main RL loss. They show that learning cumulants this way actually can help in learning the main task much faster. This phenomenon of discovering *useful* cumulants is called **discovery**.

Discovery of Cumulants: The term *discovery* denotes the automatic acquisition of cumulants that assist in the primary task, termed *useful cumulants*. Their proposed approach alternates between steps for discovery of pertinent predictive *questions* in the form of cumulant functions that are beneficial for mastering the primary task and estimation of corresponding *answers* through GVFs. The key concept underlying cumulant discovery lies in leveraging a question network with meta-learnable parameters, which allows the agent to autonomously discover questions. The main network is an RL agent's value network with additional heads for estimating predictive answers in the form of general values. The loss of the main network is composed of the RL TD loss and the TD losses of the GVFs. Veeriah et al. (2019) propose the following meta-gradient method for discovery of useful GVF:

1. perform a fixed number of updates of the main network's parameters using cumulants of the question network with frozen parameters, retaining the computational graph for the sequence of updates.

2. evaluate the RL TD loss with the latest set of main network parameters

3. backpropagate the RL loss through the differentiable graph of main network updates into the question network's parameters (via the cumulants in the GVF TD losses).

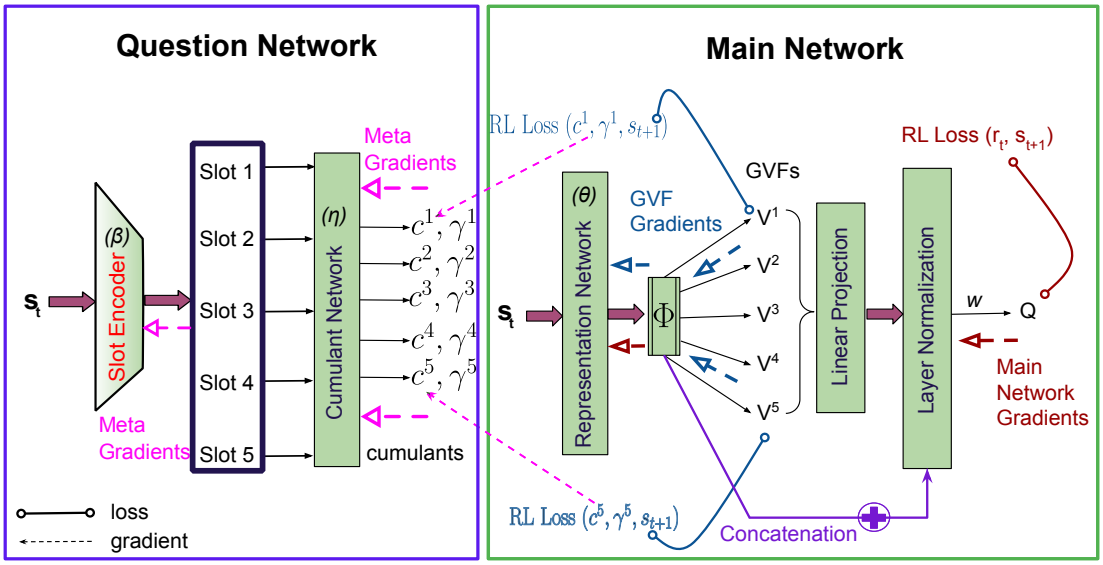


Figure 1: Sketch of the proposed architecture and gradient backpropagation pathways. The question network including the slot attention encoder gets updated with the meta-gradients from the main network. This encourages not only the cumulants to be task-specific, but slot formation is also influenced by the main RL loss, which helps slot attention focus on task-relevant features.

This approach drives the discovery of cumulants that directly minimize the task TD loss. We argue that the learned scalar GVFs can themselves be useful to generate a compact set of features that can be used to directly learn the main task. Kearney et al. (2022) extended the agent's representation with learned GVFs for control within RL. In essence, they concatenate the GVF predictions based on the learning process of the control agent and directly employ these predictions as features for enhancing the control policy.

More recently, Nath et al. (2023) developed a framework for learning cumulants in a **limited GVF regime**. They do this by attaching each GVF prediction to individual objects in the environment. Slot Attention (Locatello et al., 2020) has shown that it is possible to identify and localize objects into distinct slots within an image. This paper utilizes this architecture to tie each cumulant prediction to these generated slots. Naturally, this forces the learned GVFs to carry significantly more informative content, thereby potentially expediting performance improvements in challenging tasks.

3 Methodology

Similar to Nath et al. (2023), we use a two-network architecture, consisting of a *question network* that maps an object-centric slot representation (Locatello et al., 2020) $\left\{\text{Slot}_t^1, \ldots, \text{Slot}_t^K\right\}, \text{Slot}_t^i \in \mathbb{R}^D$ to cumulants $c^i \in \mathbb{R}$, each defining a general value function $Q_t^i = c_t^i + \gamma Q_{t+1}^i$, and a *main network* that is trained to predict the RL task value $Q_t = r_t + \gamma Q_{t+1}$ and the GVFs Q_t^i.

Question Network: The question network, depicted in Figure 1 on the left, processes batches of state observations unrolled from the replay buffer as inputs. Utilizing a slot attention mechanism, these inputs are transformed into slots, representing discovered objects from the images, each carrying object features. Slot representations are learned through forward propagation, and then mapped to GVF cumulants. This architecture is exactly similar to Nath et al. (2023), however, in that paper only the cumulant network was trained by meta-gradients, whereas the slot attention module was trained via a reconstruction objective using an autoencoder architecture. This was done in phases interspersed with meta-gradient update steps. Although using the reconstruction loss enabled the

slot attention module to capture objects well, the objects captured need not be task-specific, and thus it can capture irrelevant objects as well. We propose to instead train the slot attention module leveraging the information from the back-propagated meta-gradients used to train the question network. The entire architecture is trained by the meta-gradients obtained from the main RL loss. This ensures that the discovered slots focus more on the features relevant to the main task.

The question network consists of shared per-slot feed-forward layers that output cumulants. Given the iterative competition of randomly initialized slots in the slot attention module, the resulting slot-based representation and the resulting cumulants can be arbitrarily permuted. Similar to the set-prediction experiments in Locatello et al. (2020), we align cumulants with GVF heads of the main network by using the *Hungarian Algorithm* (Kuhn, 1955) to minimize the overall GVF loss. We observed this detail to be essential for stabilizing convergence of the model, since both the cumulants and GVF heads used in the GVF losses are updated continuously during training.

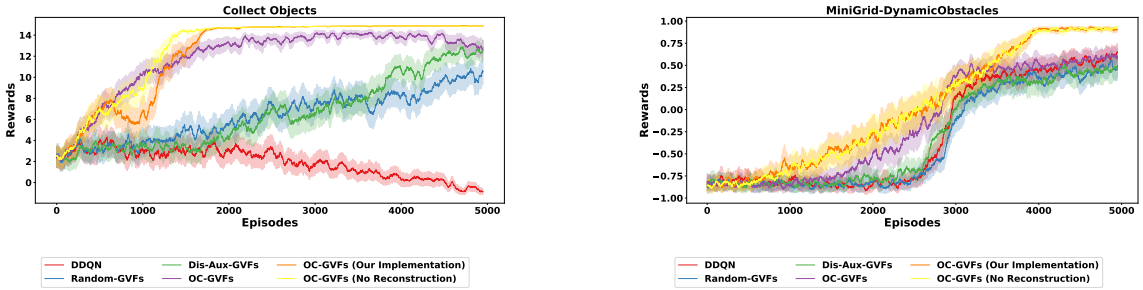
In addition to learning cumulants for each GVF, we also added discount prediction (Veeriah et al., 2019) to the architecture from Nath et al. (2023).

Main network In the architecture depicted in Figure 1, the main network, located on the right side, manages the training process for the representation, GVFs and the main RL model. In the main network, a convolutional neural network (CNN) provides a representation ϕ of the input image, followed by K GVF prediction heads. The concatenation of predicted general values V_t^i is projected, concatenated with ϕ, and then passed through layer normalization Ba et al. (2016) for stable learning. Lastly, a linear layer outputs a prediction of the RL value function.

The primary difference from Nath et al. (2023) in the Main Network is the use of the Huber TD Loss for training the GVFs instead of the mean-squared TD loss, since it is more robust to outliers. Because GVFs are trained off-policy, they can diverge more quickly. We found that using Huber Loss alleviates this problem somewhat and helps to keep the magnitude of GVFs bounded. This also allows the use of action value GVFs because of the lesser likelihood of divergent GVFs.

Training We leverage the same training protocol as Nath et al. (2023), with the following modifications. The main network parameters θ are trained via Huber TD losses of the last layer's RL value head and the Huber TD losses of the GVF heads instead of the corresponding mean-squared error terms. We also train the slot-attention module via meta-gradients instead of with a reconstruction objective.

4 Experiments



(a) Collect Objects

(b) MiniGrid Dynamic Obstacles

Figure 2: Performance of proposed algorithms in comparison to the baselines. All the algorithms are compared across 10 independent runs, with the shaded area representing standard error. These plots highlight that the meta gradient alone is enough to achieve good performance in these environments, and can sometimes work even better than reconstruction loss as in (a).

We compare performances of our proposed method — OC GVFs (no reconstruction) — with object-centric GVF (OC-GVF) (Nath et al., 2023), Discovery using Auxiliary Tasks (Dis-Aux) Veeriah et al. (2019), random GVFs and the double DQN baseline on on CollectObjects (Nath et al., 2023) and MiniGrid Dynamic Obstacles (Chevalier-Boisvert et al., 2023). We also present an adaptation of OC-GVF that uses all of our proposed changes, except that the slot-attention module is trained with reconstruction instead of meta-gradients.

From Figure 2, we observe that adding discount prediction, alignment using the Hungarian algorithm and Huber Loss has improved the performance of the original OC-GVFs (Nath et al., 2023), with our implementation showing improved policy convergence and less variance across both Collect Objects and MiniGrid Dynamic Obstacles.

5 Discussion

We study several improvements to the method for GVF discovery proposed by Nath et al. (2023). Several architectural choices make convergence more robust and allow leveraging the task-driven meta-gradient instead of an unsupervised reconstruction loss to train the slot-attention module. In our experiments, we show that our design choices improve the performance of the original method and that meta-gradients can be used to train the slot attention module. One limitation is that slots discovered via meta-gradients may be less task-agnostic than the object-centric representation learned via an unsupervised reconstruction objective, limiting the potential for transfer learning.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks, 2016.

Alexandra Kearney, Anna Koop, Johannes Günther, and Patrick M. Pilarski. What should i know? using meta-gradient descent for predictive feature discovery in a single stream of experience, 2022. URL https://arxiv.org/abs/2206.06485.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

Somjit Nath, Gopeshh Raaj Subbaraj, Khimya Khetarpal, and Samira Ebrahimi Kahou. Discovering object-centric generalized value functions from pixels. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '11, pp. 761–768, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0982657161.

Vivek Veeriah, Matteo Hessel, Zhongwen Xu, Richard Lewis, Janarthanan Rajendran, Junhyuk Oh, Hado van Hasselt, David Silver, and Satinder Singh. Discovery of useful questions as auxiliary tasks, 2019.

Zhongwen Xu, Hado P van Hasselt, and David Silver. Meta-gradient reinforcement learning. *Advances in neural information processing systems*, 31, 2018.